Uncertainties in Snowpack Simulations—Assessing the Impact of Model Structure, Parameter Choice, and Forcing Data Error on PointScale Energy Balance Snow Model Performance

Water Resources Research

RESEARCH ARTICLE
10.1029/2018WR023403

Special Section:
Advances in Remote Sensing, Measurement, and Simulation of Seasonal Snow

Uncertainties in Snowpack Simulations—Assessing the Impact of Model Structure, Parameter Choice, and Forcing Data Error on Point-Scale Energy Balance Snow Model Performance

Daniel Günther[1], Thomas Marke[1], Richard Essery[2], and Ulrich Strasser[1]

[1]Department of Geography, University of InnsbruckInnsbruck, Austria, [2]School of GoeSciences, University of Edinburgh, Edinburgh, UK

Abstract In this study, we assess the impact of forcing data errors, model structure, and parameter choices on 1-D snow simulations simultaneously within a global variance-based sensitivity analysis framework. This approach allows inclusion of interaction effects, drawing a more representative picture of the resulting sensitivities. We utilize all combinations of a multiphysics snowpack model to mirror the influence of model structure. Uncertainty ranges of model parameters and input data are extracted from the literature. We evaluate a suite of 230,000 model realizations at the snow monitoring station Kühtai (Tyrol, Austria, 1,920 m above sea level) against snow water equivalent observations. The results show throughout the course of 25 winter seasons (1991–2015) and different model performance criteria a large influence of forcing data uncertainty and its interactions on the model performance. Mean interannual total sensitivity indices are in the general order of parameter choice < model structure < forcing error, with precipitation, air temperature, and the radiative forcings controlling the variance during the accumulation period and air temperature and longwave irradiance controlling the variance during the ablation period, respectively. Model skill is highly sensitive to the snowpack liquid water transport scheme throughout the whole winter period and to albedo representation during the ablation period. We found a sufficiently long evaluation period (>10 years) is required for robust averaging. A considerable interaction effect was revealed, indicating that an improvement in the knowledge (i.e., reduction of uncertainty) of one factor alone might not necessarily improve model results.

Key Points:
- Snow models are most sensitive to input data errors then to model structure and last to model parameters
- Sensitivity analyses of complex snow models must include interaction effects for a representative assessment
- Sensitivities are variable between winter seasons, and a long period needs to be evaluated for robust averages

Supporting Information:
- Supporting Information S1

Correspondence to:
D. Günther,
daniel.guenther@uibk.ac.at

Citation:
Günther, D., Marke, T., Essery, R., & Strasser, U. (2019). Uncertainties in snowpack simulations—Assessing the impact of model structure, parameter choice, and forcing data error on point-scale energy balance snow model performance. *Water Resources Research*, *55*, 2779–2800. https://doi.org/10.1029/2018WR023403

Received 30 MAY 2018
Accepted 4 MAR 2019
Accepted article online 13 MAR 2019
Published online 5 APR 2019

1. Introduction

The water mass stored in the seasonal snowpack and the timing and intensity of the melt water release is crucial for water availability in many regions of the world and has wide implications for water resources management, impact studies, and risk assessments concerning drought and flood potential. Snowmelt models able to simulate individual physical processes rather than models mapping melt rates merely to a temperature input are advantageous for many applications. In complex situations such as rain on snow events, on climatic extremes, and for avalanche forecasting, physically based snowpack models are indispensable tools. Especially, facing a changing climate, meaningful projections of the impacts on the cold regions water cycle require robust snow models and knowledge about the associated uncertainties.

Over the last three decades, many one-dimensional models have been developed to predict seasonal snow cover on the ground. They all account for the accumulation, storage, and melt of snow. Many of them follow the same principle and solve the coupled energy and mass balance of the snowpack in a physically oriented manner. However, the abstraction of the underlying physical processes still varies in complexity. In analogy to Vionnet et al. (2012), we distinguish between three types of energy balance snow models. Model conceptualizations range from simple single-layer representations (model type 1; e.g., Strasser & Marke, 2010; Tarboton & Luce, 1996) to very detailed multilayer snow physics models (model type 3; e.g., Bartelt & Lehning, 2002; Vionnet et al., 2012), with an explicit description of the snow microstructure and its evolution over time. Unlike in type 3 models, snow layers in physically based models of medium complexity (model type 2; e.g., Marks et al., 1998) do not mimic real-world layering; rather, they are numerical constructs required to simulate vertical mass and energy fluxes in the snowpack.

Many process representations exist estimating these fluxes and various levels of complexity can be found here as well. This study focuses on type 2 models, which are widely applied in research and operational applications alike (e.g., Pomeroy et al., 2016; Winstral et al., 2013). Typically, process-based models are not calibrated, relying on their parameters being physically meaningful and determinable in the field. In reality, however, many parameters of physically based snow models are still abstract, far from being easy to identify from measurements, or observations are lacking (e.g., albedo decay time scale, and aerodynamic roughness length; Brock et al., 2006; Gromke et al., 2011; Strasser et al., 2004). Thus, understanding how parameter uncertainty propagates through snow models with different process representations (i.e., structures) is of great interest. The meteorological forcings needed to run the models are also prone to errors, especially when they have to be spatially interpolated from surrounding climate station recordings or are provided by atmospheric models. In this study, we focus on the above-mentioned uncertainty classes (i) forcing error (or input data error), (ii) model structure (or process representation), and (iii) parameter choice.

In the past, major efforts have been made to assess the linkage between model structure and overall model performance (Blöschl & Kirnbauer, 1991; Essery et al., 2013; Etchevers et al., 2004; Magnusson et al., 2015; Mosier et al., 2016), how forcing error characteristics propagate through snow models of various complexity (Lapo et al., 2015; Raleigh et al., 2015; Sauter & Obleitner, 2015), and how approximations of unmeasured forcing variables contribute to model uncertainty (Harder & Pomeroy, 2014; Marks et al., 2013; Raleigh et al., 2016).

Essery et al. (2013) compared the performance of 1,701 combinations of different snow process representations (snow compaction, fresh snow density, snow albedo, surface heat and moisture fluxes, snow cover fraction, snowpack hydraulics [liquid water transport], and thermal conductivity) with varying degrees of complexity with the help of a multimodel energy balance framework at an alpine site in France. Magnusson et al. (2015) extended the model intercomparison study using the same multimodel framework, the detailed snowpack model SNOWPACK (Bartelt & Lehning, 2002), and two simpler temperature index melt models for hydrological applications at two alpine sites. Both studies could not identify a single best model structure outperforming other models rather than a group of models with consistently good results. No clear correlation between model structure and performance could be identified above some minimum requirements. In both studies parameter values within the multimodel framework were taken from the literature and little effort has been undertaken to show the influence of parameter uncertainty in their findings. Essery et al. (2013) showed that the calibration of a previously weak performing uncalibrated model structure led to a significant improvement but not to the level of other uncalibrated model configurations. However, how robust their findings are in the face of parameter uncertainty remains unclear. Therefore, quantifying how much the model performance is attributed to the parameter choice, compared to the model structure remains an interesting question in snow modeling.

It has been hypothesized that errors in the forcing as well as in the validation data are the greatest factors affecting the model performance (Magnusson et al., 2015). Raleigh et al. (2015) investigated various scenarios of forcing errors and their propagation through snow models. These authors found a large influence of forcing biases on snowpack simulations over different output metrics and were able to identify the relative importance of individual forcing variables on snow model output variance using a sensitivity analysis (SA) framework. A comparison of the forcing error uncertainty to the model structure uncertainty originating from the 1,701 model structures presented by Essery et al. (2013) revealed that even forcing error scenarios with moderate precipitation errors have a larger influence on snowpack simulations than different model structures (for peak snow water equivalent [SWE], ablation rates, and snow disappearance) at one site.

In the hydrological modeling community, there is a growing sense for the need to systematically account for data and model uncertainties (Kavetski et al., 2006). Particularly in conceptual rainfall-runoff modeling, where parameter calibration techniques are used, accounting for various sources of uncertainty was shown to improve the robustness of the optimized parameter sets and hence improved simulation results and output uncertainty estimations (Ajami et al., 2007). SA is a common tool to investigate and apportion model output uncertainty to the different input factors and many techniques exist to quantify sensitivities (Saltelli et al., 2006). The Sobol' (1993) method is one such technique based on variance decomposition. It maps variation in the output of a numerical model to a variation of its input factors. Baroni and Tarantola (2014) showed for a soil-hydrological model how the Sobol' method could be used (in the realm of a General Probabilistic Framework) to assess the relative contribution of different sources of uncertainty related to input errors,

parameter choices, and model structure and their interaction effects. In this study, we want to extend this approach toward physically based snowpack simulations.

Uncertainty assessments in snow hydrology either focused on the impact of different process representations alone, or did not relate their findings to forcing errors, or did not investigate the robustness of their results in the face of parameter uncertainty. Therefore, the influence of forcing data error, snow model structure, and parameter choice and its interaction effects on snow model performance is yet to be investigated comprehensively. The recent literature states that "it would be interesting to assess the interplay between coexisting uncertainties in forcing errors, model parameters, and model structure, and to test how model sensitivity changes in relation to all three sources of uncertainty" (Raleigh et al., 2015). In this study, we aim to systematically investigate the influence of various uncertainty sources common for 1-D snow simulation of medium complexity (type 2 models) and assess the impact of model structure, parameter choice, and forcing data error and hence address this very research need.

We focus on comparing the impact of

a. forcing data error magnitude, parameter choice, and model structure;
b. individual forcing variables (while perturbing parameters and model structure);
c. various process representations (while perturbing forcing errors and parameters) on snow model performance.

Evaluating the model behavior during a single winter season for multiple sites was the approach taken by many intercomparison studies (e.g., Essery et al., 2013; Raleigh et al., 2015; Lapo et al., 2015). However, linking model sensitivities to environmental characteristics could not yet be demonstrated. Even for a single site the interannual variability of winter meteorological conditions (and consequently snowpack processes) might govern average model sensitivities, making it difficult to relate the findings to site characteristics, if just one winter season is evaluated. We hypothesize that (i) the evaluation time period is crucial for a representative average sensitivity assessment, (ii) the forcing error dominates output uncertainty and that the parameter choice is as important as the model structure for model performance, and (iii) neglecting interaction effects between different sources of uncertainty leads to an unrealistic quantification of relative uncertainty contributions.

The Sobol′ SA within the General Probabilistic Framework (Baroni & Tarantola, 2014) is an ideal tool to test these hypotheses, as it allows for grouping of input factors. Therefore, scalar, nonscalar, and correlated inputs can be considered. Additionally, a grouping reduces the dimensionality of the problem and hence its computational expenses. We present a series of SA designs for objectives (a), (b), and (c) and show the effects of the evaluation time period and the effect of the simultaneous assessment of multiple uncertainty sources.

2. Methods

2.1. Study Site and Data

In 1987, Kirnbauer and Blöschl (1990) installed the well-equipped snow monitoring station *Kühtai*, approximately 30 km west of Innsbruck, Tyrol, Austria (1,920 m above sea level, 47.2071°N, 11.0060°E) and initiated one of the earliest and most detailed snow studies in the Alps (Blöschl & Kirnbauer, 1991, 1992; Blöschl, 1991; Blöschl, Gutknecht, & Kirnbauer, 1991; Blöschl, Kirnbauer, & Gutknecht, 1991). Measurements and station maintenance have continued ever since. Twenty-five years of this novel data set (1990–2015) are freely available and include quality controlled time series of meteorological observations and snow physical properties (Parajka, 2017; Table 1). While all meteorological variables have been gap filled to some extent, wind speed measurements are lacking from 2000 onward and have been constructed from surrounding station recordings using established correlations during the time of operation (Krajči et al., 2017). The snow monitoring station Kühtai offers a suite of snowpack observations, including a hexagonal 10-m^2 snow pillow for SWE observations. Operation of the pillow failed in the water years 1996, 2013, and the latter half of 2012, resulting in a validation data set of 22 and 23 years for ablation and accumulation, respectively. A map of the study location is provided in the supporting information (Figure S1), and pictures of the snow monitoring station can be found in Krajči et al. (2017).

In Figure 1, we present some meteorological characteristics of the analyzed data set. The study site is characterized by average winter air temperatures of −2.65 °C (November to April), with 2007 being the warmest (−0.17 °C) and 2006 being the coldest (−4.5 °C) winter season. Mean annual precipitation is about 1,150

Table 1
Forcing and Validation Data Observations at the Kühtai Station

Variable	Symbol	Period of operation	Filled gaps (1990–2015; %)
Air temperature	T_{air}	1990–2015	0.03
Precipitation	P	1990–[a], 2001–2015[b]	1.4
Incoming shortwave radiation	Q_{si}	1990–2015	11.5
Relative humidity	RH	1990–2015	1.6
Wind speed	U	1990–1999	70
Snow water equivalent	SWE	1990–2015	Not filled

[a] Self-cast tipping bucket. [b]OTT Pluvio.

mm, while SWE observations peak at about 375 mm in average. Typically, the snowpack lasts from November to May, with only 4 mm of midwinter melt (December–February). However, melt events just before peak SWE are common in March and April resulting in an average of 30-mm melt (1 March until observed peak SWE). Melt water release between 1 October and observed peak SWE range between 16.6 mm in 1992 and 137.7 mm in 2003.

2.2. Estimation of Unmeasured Forcing Data

Meteorological observations not recorded but needed to force snow simulations include longwave irradiance and information about the precipitation phase. Time series of incoming longwave radiation is calculated offline using the meteorological preprocessor of the hydroclimatological model AMUNDSEN (Strasser, 2008, 2008) coupled to a spatially distributed version of the Factorial Snowpack Model (FSM; Essery, 2015). Total longwave irradiance is calculated as the sum of longwave radiation emitted from the clear-sky atmosphere, the cloud cover, and the surrounding terrain. During daytime, cloud coverage was estimated by relating potential incoming solar radiation to actual recordings. Potential incoming clear-sky shortwave radiation is calculated following Corripio (2002), taking into account transmission losses due to scattering and absorption, multiple reflections between the atmosphere and the surface, and reflections and shading by the surrounding topography. Atmospheric transmissivity (τ_{atm}) can then be computed as the ratio of actual and potential global radiation. Assuming its transferability for similar mountainous topographies we invert a fit function found by Greuell et al. (1997) on Pasterze glacier (Austria) relating τ_{atm} and the cloud fraction (cn).

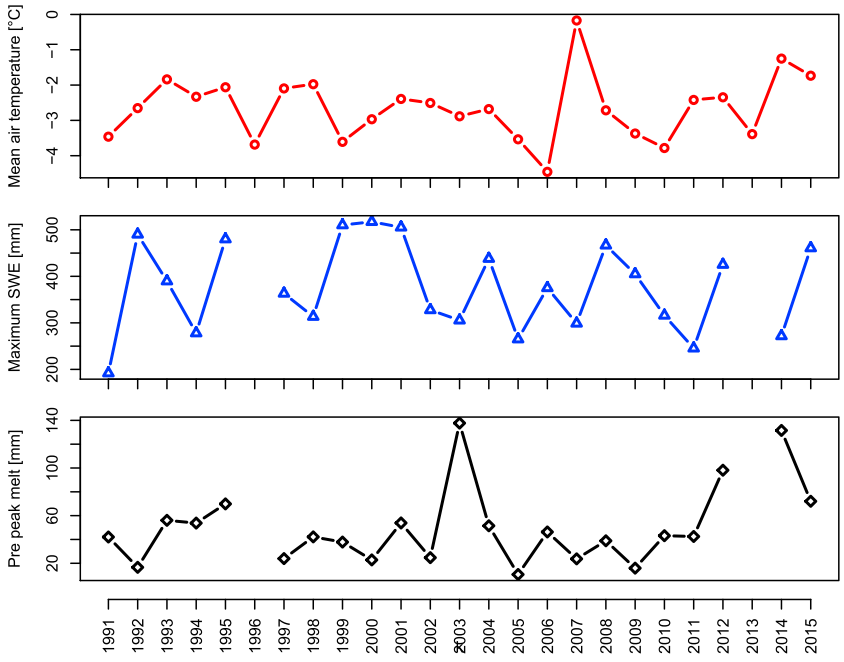

Figure 1. Mean air temperatures from November to April (top), maximum snow water equivalent (middle), and snowmelt before observed peak snow water equivalent (SWE; lower). Tick marks represent water years.

Table 2
Process Representations Available in the Factorial Snowpack Model (Essery, 2015)

Process	Implementation	Option
Absorption of solar radiation	Albedo evolution	0: function of surface temperature 1: decays with time
Heat conduction in snow	Thermal conductivity	0: constant 1: function of snow density
Compaction of snow	Snow density	0: constant 1: compaction
Transfer of heat to snow from the air	Correction for atmospheric stability	0: off 1: on
Transport of liquid water	Snowpack hydraulics	0: immediate drainage 1: bucket model

During nighttime, with no shortwave radiation measurements available, the cn is calculated following Liston and Elder (2006), relating cn to the humidity expanded to the 700-mb level. Clear-sky emissivity (ε_{cs}) is approximated following Klok and Oerlemans (2002), while the emissivity of overcast skies (ε_{oc}) is assumed to be 0.975 as in Greuell et al. (1997). Now the all-sky emissivity (ε_{sky}) can be computed as

$$\varepsilon_{sky} = \varepsilon_{cs} \cdot (1 - cn^2) + \varepsilon_{oc} \cdot cn^2. \tag{1}$$

Subsequently, longwave irradiance from the fraction of the visible sky and the surrounding slopes is calculated following the Stefan Boltzmann law. Assuming that the surrounding bare or snow-covered slopes emit as a black body in the infrared spectrum, outgoing longwave radiation from these surfaces is calculated depending on their surface temperature.

Precipitation phase partitioning based on near-ground meteorological observations is a rather uncertain, but standard practice in hydrology (Harder & Pomeroy, 2013). Hereby, a common challenge is to find suitable parameter sets required by most methods. Marked spatial variation in threshold air temperature across the Northern Hemisphere could be identified recently (Jennings et al., 2018). Even though many approaches exist to relate snowfall fraction to air temperature (T_{air}) and humidity (RH) via various formulations, preliminary investigations indicate that the uncertainty introduced through the parameter choice is much larger than the effect of the applied phase determination method (Günther et al., 2017). Linking the phase decision to one single transition temperature is advantageous in the presented analysis design, as it allows treating the parameter analogously to other input data errors. Due to computational considerations, we relate the snowfall fraction (f_r) to an air temperature threshold (T_{ph}), leading to a binary prediction of rainfall and snowfall.

$$f_r(T_a) = \begin{cases} 1, & T_{air} < T_{ph} \\ 0, & T_{air} \geq T_{ph} \end{cases} \tag{2}$$

2.3. Snow Model and Validation Metrics

The open-source FSM (Essery, 2015) solves the coupled mass and energy balance of a snowpack in a control volume of 1-m^2 surface area and height H_s. FSM is a type 2 snow model with a user-selected maximum of three snow layers. Total snow depth governs the number and thickness of the snow layers. The model was developed to allow for systematic investigations of the interplay between different snowpack process representations. It offers the possibility to choose from two different representations of each of the following processes: absorption of shortwave radiation, heat transfer in snow, densification of the snowpack, turbulent transfer of energy, and liquid water storage in the snowpack (Table 2). For each process FSM allows a simpler representation (option 0) and a more complex/prognostic one (option 1). Model structure and available process combinations are explained in detail in Essery (2015). Physically based energy balance models allow for detailed but efficient simulations of coupled snow processes. The comparatively short run time and the rather small set of model parameters make FSM ideal to investigate various model configuration and parameter settings.

The simulated daily snowpack predictions are validated against snow pillow SWE observations. Model skill is determined using a suite of performance criteria. We present three model performance metrics, which quantify the model ability to predict daily SWE observations as (i) the mean absolute error (MAE) during the full winter season (hereafter "full season"), (ii) the MAE during the main accumulation period (from 1 October until the maximum SWE recordings, hereafter "accumulation"), and (iii) MAE of snow mass changes during the main ablation period (from observed peak SWE onward, hereafter "ablation"). To compare model performances between years of different snow season length and to avoid overvalueing the ability of the model to predict snow-free conditions, we only evaluate during periods when either an observed or simulated snow cover is present. Additional performance metrics were explored and presented in the supporting information document. They include the Kling-Gupta model efficiency (KGE) (Gupta et al., 2009) for the full winter season, MAE of positive snow mass changes, MAE of negative snow mass changes, and errors in ablation slope, snow cover duration, 1 April SWE, peak SWE, and timing of peak SWE.

We analyze different performance criteria, focusing on different snowpack system states, as it is expected that the sensitivities vary during the winter season (Sauter & Obleitner, 2015). During the accumulation period the sunshine hours and solar angles are low and the surface albedo is generally high, resulting in a limited input of solar energy. During this period, we do not expect any specific energy flux to be the dominant source of energy. When the snowpack is cold (i.e., if a cold content is still present) energy is not contributing to melt, and hence, SWE is rather insensitive to errors in the energy balance. Later in the winter season, with rising air temperatures and solar angles, snowpack temperatures increase (depleting the cold content). Once the snowpack reaches an isothermal state (at 0 °C), SWE becomes sensitive to errors in the energy fluxes, as they translate directly to the energy available for melt.

In the realm of the presented SA the computed model performances are the final outputs of the system. Hence, we compute the influence of different sources of uncertainty not on predicted snow mass but on the model skill itself.

2.4. Sensitivity Analysis and the General Probabilistic Framework

The sensitivity of the model system to a change in input data, parameter choice, and model structure is quantified using the Sobol' SA framework, a global variance-based method (Sobol, 1993). The variance in the model prediction (Y) resulting from a change of one specific parameter (X_i) is described as the first-order sensitivity index (S_i). The total-order sensitivity index (S_{Ti}) further includes the interaction effect of all other parameters. S_i and S_{Ti} are estimated as

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)}, \tag{3}$$

$$S_{Ti} = 1 - \frac{V[E(Y|X_{\sim i})]}{V(Y)}, \tag{4}$$

where V is the variance operator, E is the exception operator, and $X_{\sim i}$ denotes all parameters except X_i. Both sensitivity indices range from 0 to 1, denoting the fraction of the output variance explained. However, the sum of all S_{Ti} values can exceed one, since interaction effects are credited to all variables involved. In order to decompose the variance and for evaluation of equations (3) and (4) a triplet of sampling matrices A, B, and $A_B^{(i)}$ is generated. Matrices A and B consist of $N \times k$ (number of samples × number of parameters) quasi-random samples in the range [0,1]. $A_B^{(i)}$ is then constructed from a combination of A and B, taking all columns from A except the ith column, which is taken from B (Saltelli et al., 2010). This results in a matrix of dimension $kN \times k$. Following Jansen (1999) and Saltelli et al. (2010), first-order and total-order sensitivity indices are calculated as

$$S_i = 1 - \frac{\frac{1}{2N}\sum_{j=1}^{N}(f(B)_j - f(A_B^{(i)})_j)^2}{V(Y)}, \tag{5}$$

$$S_{Ti} = \frac{\frac{1}{2N}\sum_{j=1}^{N}(f(A)_j - f(A_B^{(i)})_j)^2}{V(Y)}. \tag{6}$$

Table 3
Uncertainty Sources and Their Distributions: Input

Source	Description	Unit	Distribution	Distribution parameter (1st and 99th percentiles)	Reference
T_a	Air temperature	°C	Normal	$\mu = 0, \sigma = 1.3\ (-3, +3)$[a]	Raleigh et al. (2015)
RH	Relative humidity	%	Normal	$\mu = 0, \sigma = 10.7\ (-25, +25)$[a]	Raleigh et al. (2015)
U	Wind speed	m/s	Normal	$\mu = 0, \sigma = 1.3\ (-3, +3)$[a]	Raleigh et al. (2015)
Q_{si}	Incoming shortwave radiation	W/m^2	Normal	$\mu = 0, \sigma = 43\ (-100, +100)$[a]	Raleigh et al. (2015)
Q_{li}	Longwave irradiance	W/m^2	Normal	$\mu = 0, \sigma = 10.7\ (-25, +25)$[a]	Raleigh et al. (2015)
P	Precipitation	—	Normal	$\mu = 1.05, \sigma = 0.107\ (0.8, 1.3)$[b]	Station recordings[c]
T_{ph}	Phase transition air temperature	°C	Uniform	$\min = -1, \max = 2.5$[a]	Ye et al. (2013)

[a] Additive. [b]Multiplicative. [c]Relative differences of annual precipitation sums from automated and manual recordings.

Simultaneous evaluation of S_i and S_{Ti} requires $N(k + 2)$ simulations. Convergence is tested by means of bootstrapping with replacement. We report the standard error (SE) of 1,000 samples. In this study, we want to quantify the effect of 7 input variables, 5 snow process representations, and 7 to 14 model parameters (depending on the configuration) on snow model performance. Resolving the sensitivity indices of every single factor is both computationally not feasible and very hard to interpret, since the total-order sensitivity indices will include interaction effects from potentially all other factors. Therefore, we follow the concept of the General Probabilistic Framework (Baroni & Tarantola, 2014) and group uncertainty sources into wider classes according to the research aims defined in section 1. Now for each specific class, n-independent realizations are generated to mirror the uncertainty inherent to that class. Each realization is associated with an integer number in the range $[0, n]$ from which later samples are drawn in the course of the SA. Details about this procedure are given in section 2.5.

As mentioned, the sensitivities of single parameters are not investigated in this study. We strive to analyze the impact of various uncertainty sources simultaneously over multiple winter seasons. This enables us to include and quantify interaction effects and hence get a more representative picture of model sensitivity. Careful analysis of the findings increases our understanding of what governs model performance and future work will help to identify robust model settings, applicable in climate change studies, or suitable for ungauged basins. Evaluating individual model realizations and identifying well-performing combinations is not within the scope of this study.

2.5. Workflow

2.5.1. Step 1: Define the Sources of Uncertainty and Their Distributions

Assessing the impact an uncertain variable has on a system requires (i) identifying the variable of interest and (ii) information about its uncertainty distribution. Given that all parts in the modeling chain are subject to some degree of uncertainty, it seems arbitrary to identify single sources of uncertainty. For these tasks, we therefore follow previous studies as close as possible to ensure comparability.

Input data error: Raleigh et al. (2015) explored the impact of different forcing error characteristics on snow simulations, employing, that is, uniform (*UB*) and normally distributed biases (*NB*) of different magnitudes. We follow the authors NB scenario, except for the precipitation (P) error, introducing a bias to measured forcing variables. Whereas the authors published value ranges for normally distributed T_{air}, wind speed (U), relative humidity (RH), shortwave (Q_{si}), and longwave (Q_{li}) irradiance errors, in this study, we fit normal distributions, reproducing the "range" values at the 1st and 99th percentiles. Precipitation biases are obtained as a normal distribution derived from relative differences of annual precipitation sums from automated and manual recordings at the Kühtai station. We assume that this procedure mimics typical P errors introduced by wind undercatch at the presented site. Uncertainty originating from the approximation of the snowfall fraction is reflected via a uniform distribution of the transition air temperature. Distribution parameters of input data errors are displayed in Table 3. Resulting 1st and 99th quantiles are given for context.

Parameter choice: Depending on the snow model configuration, 7 to 14 parameters are selected for the SA. Four of these parameters are used by all FSM configurations (α_{max}, α_{min}, h_f, and z_{0s}); all other parameters depend on the individual process options selected. Simpler options feature fewer parameters, the more complex prognostic options require more. We describe the uncertainty originating from individual parameters as uniform distributions ranging between values documented in the literature, if possible (Table 4), to restrict the analysis to a physically meaningful parameter space.

Water Resources Research

10.1029/2018WR023403

Table 4
Uncertainty Sources and Their Distributions: Parameter

Source	Unit	Description	Distribution	Distribution parameter	Reference
α_{max}	—	Maximum albedo for fresh snow	Uniform	min = 0.75, max = 0.95	Singh and Singh (2001)
α_{min}	—	Minimum albedo for aged snow	Uniform	min = 0.5, max = 0.6	Pomeroy and Brun (2001)
b_h	—	Atmospheric stability adjustment parameter	Uniform	min = 4, max = 6	Louis et al. (1982) $b = 5$
b_λ	—	Thermal conductivity exponent	Uniform	min = 1.9, max = 2.1	No reference
h_f	m	Snow cover fraction depth scale	Uniform	min = 0.05, max = 0.2	No reference
λ_0	$W{\cdot}m^{-1}{\cdot}K^{-1}$	Fixed thermal conductivity	Uniform	min = 0.105, max = 0.699	Sturm et al. (2002)
ρ_0	kg/m^3	Fixed snow density	Uniform	min = 208, max = 306	Range of mean annual density recordings
ρ_f	kg/m^3	Fresh snow density	Uniform	min = 70, max = 190	Cline et al. (1998)
ρ_{cold}	kg/m^3	Maximum density for cold snow	Uniform	min = 200, max = 400	Singh and Singh (2001)
ρ_{melt}	kg/m^3	Maximum density for melting snow	Uniform	min = 400, max = 650	Singh and Singh (2001)
S_α	kg/m^2	Snowfall required to refresh albedo	Uniform	min = 1, max = 10	Wang et al. (2013); Douville et al. (1995)
T_α	°C	Albedo decay temperature threshold	Uniform	min = −4, max = −0.1	No reference
τ_{cold}	hr	Cold snow albedo decay timescale	Uniform	min = 750, max = 1,250	No reference
τ_{melt}	hr	Melting snow albedo decay timescale	Uniform	min = 50, max = 150	No reference
τ_ρ	hr	Compaction timescale	Uniform	min = 100, max = 300	No reference
W_{irr}	%	Irreducible liquid water content	Uniform	min = 3, max = 8	Singh and Singh (2001)
z_{0s}	m	Roughness length of snow-covered ground	Uniform	min = 0.001, max = 0.0168	Helgason and Pomeroy (2011); Daniel Moore (1983)

Model structure: As specified in section 2.3 and Table 2, model uncertainty is approximated via 32 different FSM configurations. The ensemble spread generated by these 32 combinations was found to be similar to previous model intercomparisons with a larger ensemble (1,701) but with much shorter run times (Essery, 2015).

2.5.2. Step 2: Group Into Wider Uncertainty Classes

Investigating the influence of multiple uncertainty sources on snow model performance requires an appropriate classification of model elements. Attributing a model element to a specific uncertainty class is often a fuzzy choice and can be blurred by the specific location of a calculation in the modeling chain. For instance, in contrast to FSM, some snow models take the overall precipitation sum as an input and compute rain and snowfall volumes internally (e.g., Strasser & Marke, 2010). Another example is the prerequisite of some snow models to preprocess net shortwave radiation, forcing the user to compute surface albedo off-line (e.g., Marks et al., 1998). Additionally, the approximation of unmeasured forcing data like precipitation phase and longwave irradiance relies on an estimation method (a model) with an inherent structural uncertainty and parameter uncertainty. Hence, a clear classification might be model and data dependent.

In this study, we address these ambiguities by relating the uncertainty classification to processes above and below the snow-atmosphere interface. Model structure uncertainty is seen as the uncertainty due to different representations of snowpack processes. The parameter choice uncertainty relates to the uncertainty of the corresponding parameter value of these snowpack process representations. For example, the process representations used to compute surface albedo, and hence, net shortwave radiation is grouped into the "model structure" class, as this property is mostly dependent on snow surface processes. The parameter values used in the respective representation is grouped in the "parameter choice" class. Following this logic the determination of the precipitation phase, is grouped into the "input data error" class, as the transition of rain and snowfall is the result of atmospheric processes, in spite of the fact that its estimation relies on a model itself. Analogously, we do not investigate the structural and parameter uncertainty of the longwave irradiance estimation and treat this variable as it was measured in the input data error class. We acknowledge that this classification scheme still lacks a certain degree of conceptual clarity, since we neglect any interplay between the snow surface and the meteorological variables (e.g., longwave emission and shortwave reflectance from surrounding slopes, influence of the snow surface on near-ground measurements of wind speed, temperature, and humidity).

The three sensitivity analyses presented in this study aim to compare the influence of forcing data error, model structure, and parameter choice (objective a), compare the sensitivities to individual forcing variables (objective b), and compare the impact of different snowpack process representations (objective c). Hence, three different grouping designs are required. The 29 individual sources of uncertainty shown in step 1 are grouped into wider uncertainty classes.

a. The uncertainty associated with the six meteorologic forcing variables, as well as the precipitation phase transition parameter $_T$ph, are grouped into the uncertainty class input data error. All possible combinations of FSM process representations are pooled into the model structure class. Snow model parameters (section 2.5.1) are summarized into the group parameter choice.
b. To resolve the impact of all seven input forcings individually, these uncertainty sources are not further grouped. However, in order to include interaction effects with other parts of the modeling chain, all remaining uncertainty sources are pooled into one single class.
c. Analogously, the impact of all five snow model processes is resolved separately as well. Here all other uncertainty sources of "input data" and parameter choice are again grouped into one single class.

This grouping reduces the degrees of freedom in the SA from 29 to 3, 8, and 6 for the study aims (a), (b), and (c), respectively.

2.5.3. Step 3: Generate Model Realizations

For each of the uncertainty classes input data error and parameter choice, where individual uncertainty sources are lumped together, independent realizations are generated. Similar to the SA sampling, we generate a suite of n quasi-random realizations dependent on the number of parameters k as $n = N(k + 2)$, where N is set to 100. We assume that the resulting 900 and 1,900 realizations, respectively, are sufficient to represent the uncertainty associated with the classes. The uncertainty class model structure consists of $n = 2^5 = 32$ FSM option combinations. The realizations are linked to an integer number (i.e., the row number of the array).

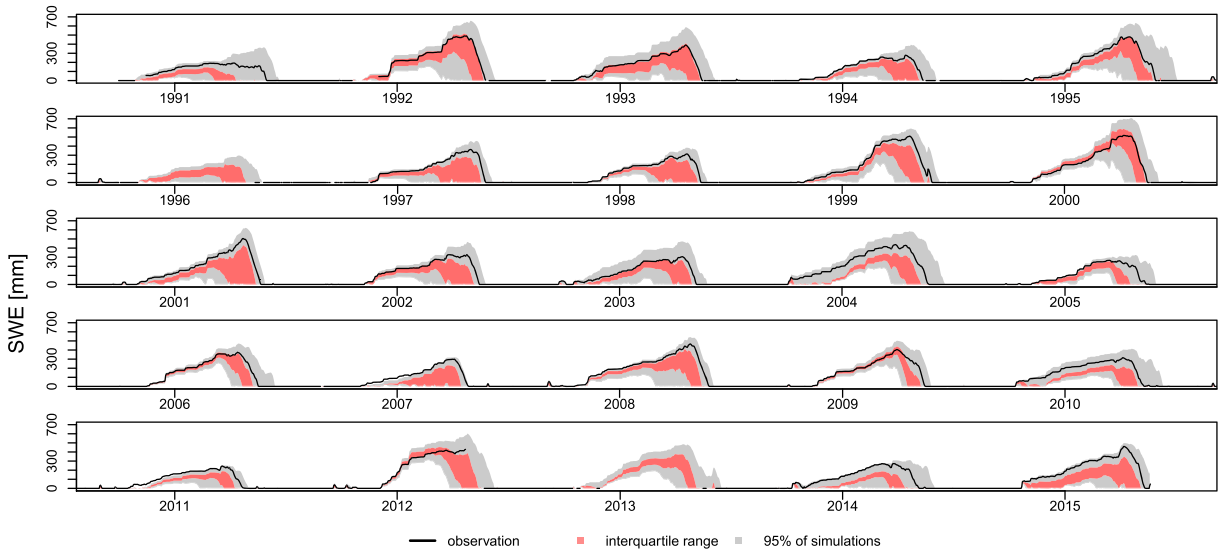

Figure 2. Ensemble spread of 50,000 snowpack simulations over the course of 25 winter seasons (objective a). SWE = snow water equivalent.

2.5.4. Step 4: Take Samples

Sampling matrices A and B are generated from $N \times k$ quasi-random samples, with $N = 10,000$. Depending on whether single uncertain parameters or previously generated classes are to be distinguished in the SA, samples are either drawn from the probability distributions specified in step 1 or from a uniform distribution of discrete factors linked to the prior generated realizations. Consequently, the three research aims (a)–(c) result in 5.0×10^4, 1.0×10^5, and 8.0×10^4 model simulations.

2.5.5. Step 5: Run the Model and Calculate Model Performance

Input data errors are introduced, model parameters are set, and the FSM options are chosen. We run the model for all generated settings over the course of 25 consecutive years, resulting in a total simulated period of 5.7×10^6 years. For each winter season (when observation data is available), we compute the model skill using various model performance criteria, including MAE during the full season and for the accumulation and ablation period separately (section 2.3). Due to the efficiency of FSM, the computational costs of snow cover simulations and subsequent performance calculations are comparatively low. In parallel processing on a standard recent desktop machine, this sums up to only 0.014 s/year, making extensive Monte Carlo simulations an overnight procedure.

2.5.6. Step 6: Calculate Sensitivity Indices

In order to ensure interpretable results, simulations predicting unrealistic amounts of snow are excluded from further analysis. Acceptable simulations include a minimum peak SWE of 10 mm and snow-free conditions at least once in the summer (no "glacier formation"). Predictions not meeting one criterion in any of the simulated years are omitted. Main and total-order sensitivity indices are calculated for each year and each performance metrics. Finally, the accuracy of the estimates is tested via bootstrapping.

3. Results

3.1. Ensemble Spread and Variability in Model Performance

Before evaluating the calculated sensitivity indices the ensemble spread is investigated (Figure 2). Of all simulations, 95% are displayed by the gray bands, and red bands show the interquartile range (i.e., 50%). When visually comparing the ensemble spread of individual years, Figure 2 reveals that there is a considerable interannual variability. In some winter seasons model predictions differ substantially (e.g., 1993, 1997, and 2007), while in other years simulation results agree much more (e.g., 2000, 2006, and 2012) and hence form a narrower band. In the first months of the water year (October–January), SWE simulations show a comparatively small spread for many winter seasons (e.g., the cold winter season 2006) until different model realizations start to diverge. For most winter seasons simulation spread encompasses observations. However, early in the winter seasons 2007, 2008, and 2011, none of the model realizations reproduces the observation at times. Just 0.25 to 1% of the model realizations (487, 356, and 200) produced unrealistic amounts of snow (see section 2.5 step 7) and are excluded from the analysis for objectives (a)–(c), respectively.

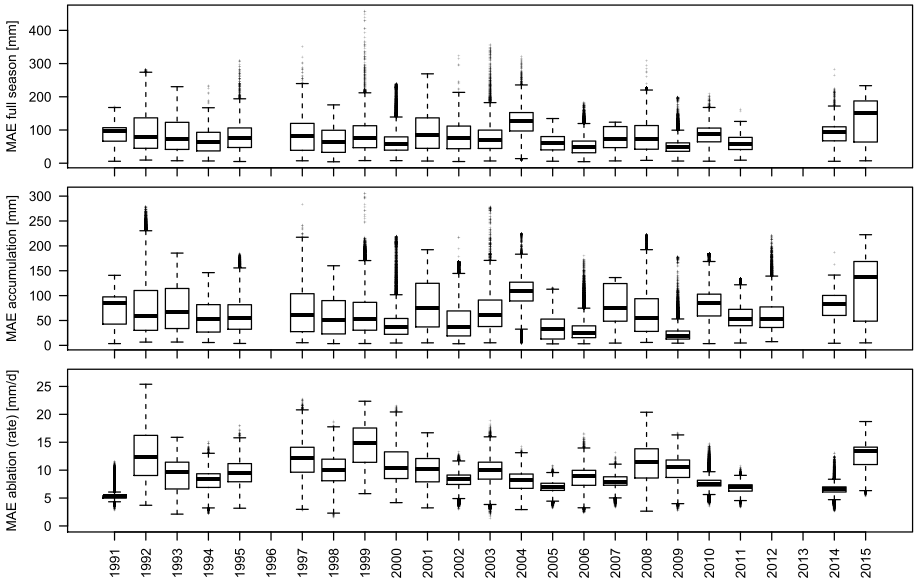

Figure 3. Simulation performances of the ensemble (objective a) for each winter season (labeled as the corresponding water year). MAE = mean absolute error.

Each simulation is validated using daily SWE observations according to section 2.3. In Figure 3, we present annual errors for the full winter season, the accumulation period, and the ablation period, respectively. Error distributions of additional performance metrics are provided in the supporting information document (Figure S2). MAE during the full winter season vary substantially from near 0 up to 450 mm. Median errors of the ensemble show the smallest values in 2006 and 2009 and the highest in 2004 and 2015. During water year 2015, we also see the highest variation in model performance. In the accumulation period, we see a similar pattern of MAE, with most model realization producing weak performance measures in 2009 and 2015, whereas errors in 2006 and 2009 are low. After peak SWE (during the ablation period) median MAE are lowest in 1991 and 2014 and highest in 1999 and 2015.

3.2. Impact of Various Uncertainty Sources

Convergence of $10,000(k + 2)$ model realizations is tested for all performed SA via bootstrapping with replacement. Of all bootstrapped SEs of main and total-order sensitivity indices, 90% were found to be smaller than 0.012 and none above 0.016 (Figure 4). These small errors show the robustness of the estimated indices and suggest convergence.

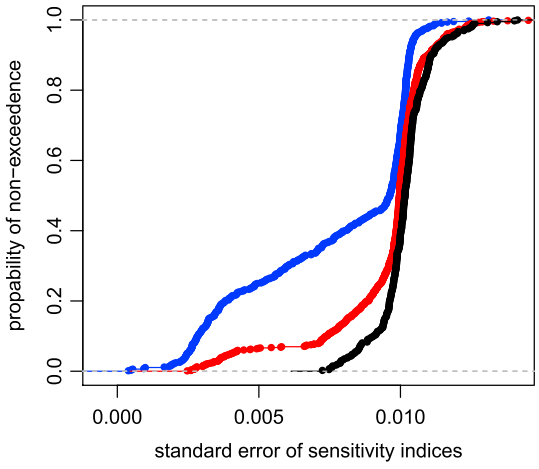

Figure 4. Empirical cumulative distribution function for bootstrapped standard errors of all calculated main and total effect sensitivity indices for objective (a; black line), (b; red line), and (c; blue line).

3.2.1. Impact of Forcing Data Error, Parameter Choice, and Model Structure (Objective a)

In objective (a), the influence of uncertainties originating from the input data, the parameter choice, and the model structure on snow model performance are compared, regarding their first-order (i.e., main) effect and their total-order (i.e., main plus interaction) effect (Figure 5). When comparing S_i and S_{Ti}, it can be seen that existing interaction effects explain a large proportion of the variance. During the full winter period, model skill is mostly sensitive to errors in the forcing data with the highest values of main and total effect indices, followed by the model structure and the parameterization. While this general average ordering of parameter choice < model structure < input data error is also preserved when evaluating just during accumulation and ablation periods, the difference between the three uncertainty classes become slightly smaller in the latter one. Although the total effect of the input error explains a statistically significant (paired Wilcoxon signed-rank test, $p < 0.001$) higher fraction of the variance during the ablation period compared to the accumulation period, the main effect is significantly reduced ($p < 0.01$). Hence,

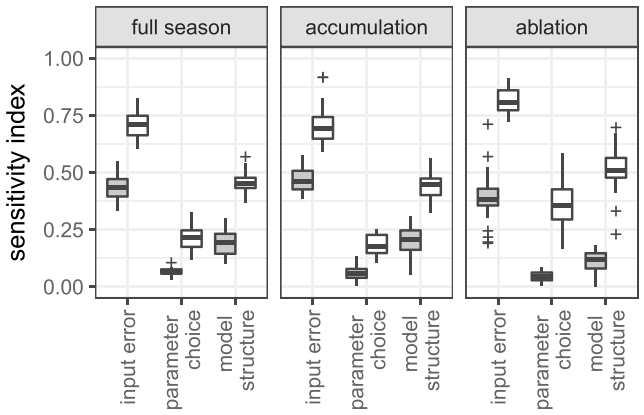

Figure 5. First (gray boxplots) and total-order (white boxplots) sensitivity indices over 23 years for the impact of input error, parameter choice, and model structure on the model skill during the whole winter season (left), the accumulation period (middle), and the ablation period (right).

a larger proportion of the variance is explained by the interaction effect. This increase in the interaction effect toward the second part of the winter season is also evident for the other two uncertainty classes parameter choice and model structure. Median total effect indices increase for parameter choice from 0.18 to 0.36 and for model structure from 0.45 to 0.51, respectively, while main effects decrease. Including nine additional model performance criteria in the SA confirmed the general ranking with two exceptions (supporting information Figure S3). We found that predicting the timing of the annual maximum SWE is as sensitive to the model structure as to the input error and that simulating correct mean seasonal ablation rates (ablation slope) is more sensitive to model structure than to input errors ($p < 0.001$).

3.2.2. Impact of Forcing Error Magnitude (Objective b)

Taking a closer look into the influence of individual forcing variables on model performance in the context of various snow model structures and parameter sets (Figure 6) reveals marked differences between the winter periods. During the full winter seasons the variance of simulation performances is mainly explained by the forcing variables precipitation, longwave irradiance, incoming shortwave radiation, and air temperature and all their interactions. The total effect of Q_{li} and T_{air} are significantly larger ($p < 0.005$) than Q_{si}. However, there is neither a statistically difference between the effect of P and Q_{si} nor between P, Q_{li}, and T_{air} (all $p > 0.2$). During the accumulation period the large interannual variability in both S_i and S_{Ti} values becomes evident for P (i.e., size of the boxplot). The total explanatory power of P ranges from 11.8% in 2015 to 84.3% in 2009. However, median sensitivity indices are not statistically different from those of Q_{si}, Q_{li}, and T_{air}. The phase transition temperature, RH and U have a significant smaller effect of accumulation model skill. Ablation performances are most sensitive to Q_{li} and T_{air}, with median values of 0.37 for both. Median PS_i and S_{Ti} values decrease significantly from accumulation period to ablation period (both $p < 0.001$) and P explains only 9% (median) of the variance in total. Also, the variability in total-order sensitivity indices is markedly reduced for P. The importance of the phase determination increases when validating after peak SWE, explaining a noticeably

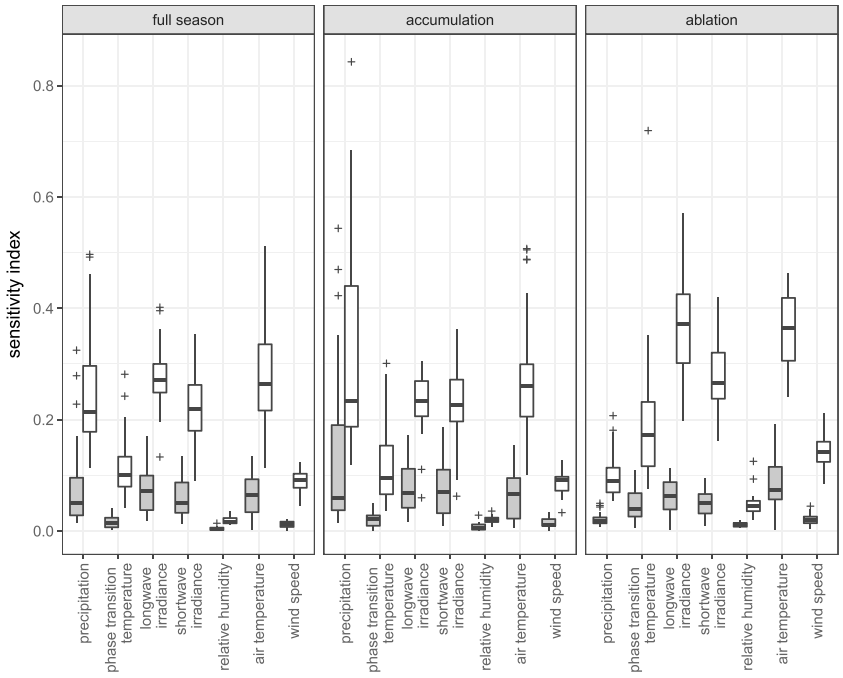

Figure 6. First (gray boxplots) and total-order (white boxplots) sensitivity indices over 23 years for the impact of various forcing errors on the model skill during the whole winter season (left), the accumulation period (middle), and the ablation period (right).

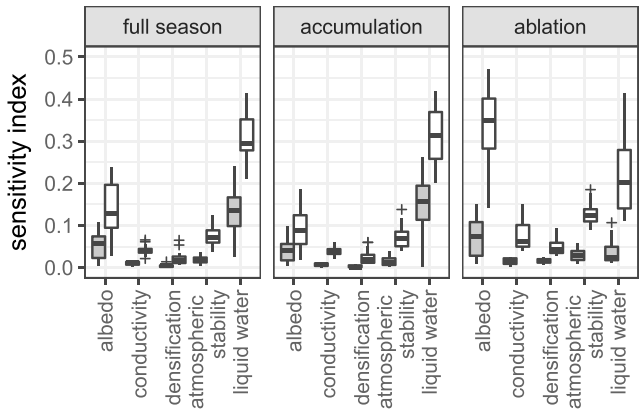

Figure 7. First (gray boxplots) and total-order (white boxplots) sensitivity indices over 23 years for the impact of various model options on the model skill during the whole winter season (left), the accumulation period (middle), and the ablation period (right).

higher fraction of the model skill variance than P. Again, the increase in interaction effects becomes apparent during the ablation periods. The influence of individual forcing errors on additional validation metrics can be found in the supporting information (Figure S4). Q_{li} has the highest impact on predicting seasonal ablation slopes and ablation timing, MAE of negative SWE changes and errors of 1 April SWE values.

3.2.3. Impact of Model Structure (Objective c)

In Figure 7, the S_i and S_{Ti} values are presented for the five model options available in FSM. For the model skill over the full winter seasons, the snowpack hydraulics option, the albedo option, and the correction for atmospheric stability have the highest impact (in this order). Evaluation only during the accumulation periods reveals the same three options as influential; however, the total albedo effect is reduced notably and differences to the stability option become less significant ($p = 0.045$). This observation stands in contrast to the sensitivities obtained through evaluation during the ablation period. Here the albedo option and its interactions explains the largest portion of the variance (median value of 35%) of all model options. While main and total effect of the liquid water transport option decrease between the accumulation and ablation period, the impact of the conductivity option, the densification option and the atmospheric stability option increase as well.

3.3. Implications of Evaluation Time

To assess the impact the number of evaluated winter seasons has on mean sensitivity estimations, an error of the mean sensitivity values is calculated for an increasing number of winter seasons. Extraction of random subsets of annual S_{Ti} values (from 1 to 22 winter seasons) are replicated 500 times, and errors are computed against the full period. Mean errors for different sample sizes are reported in Figures 8a–8c). In order to put the errors in mean total-order sensitivities into context, we compare them against the SEs of the 22-year

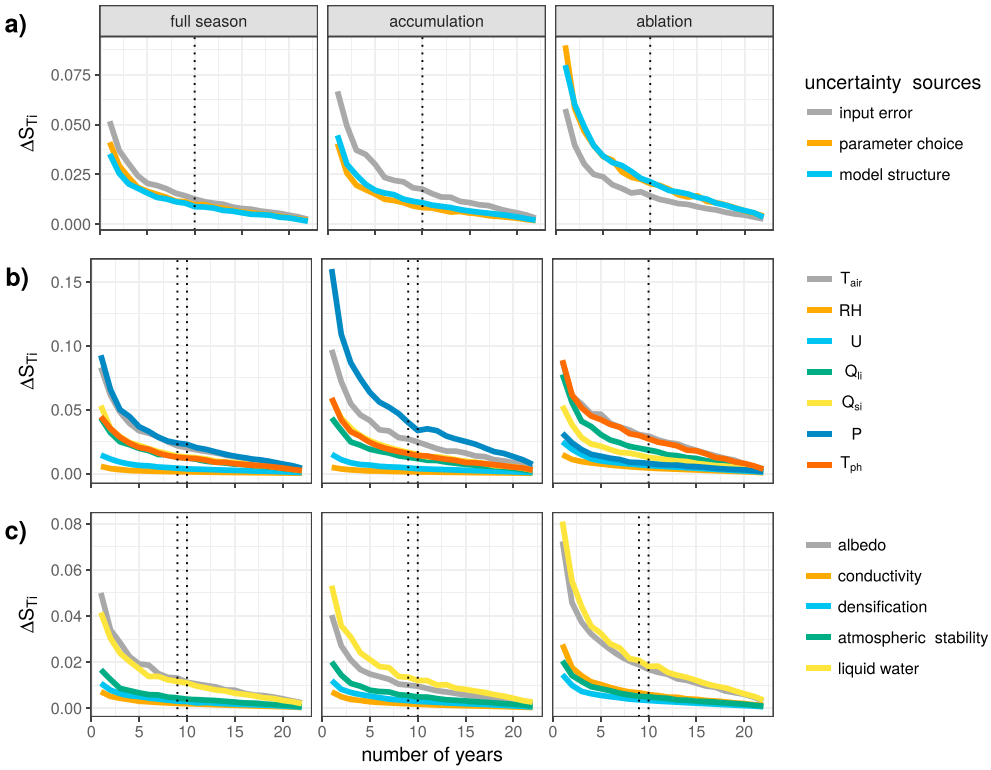

Figure 8. Mean S_{Ti} estimation errors for an increasing number of winter seasons and different uncertainty sources. (a–c) The objectives (analyzed in sections 3.2.1, 3.2.2, and 3.2.3). Dotted lines represent the number of years necessary to obtain smaller sensitivity errors than standard errors of the mean over the whole period.

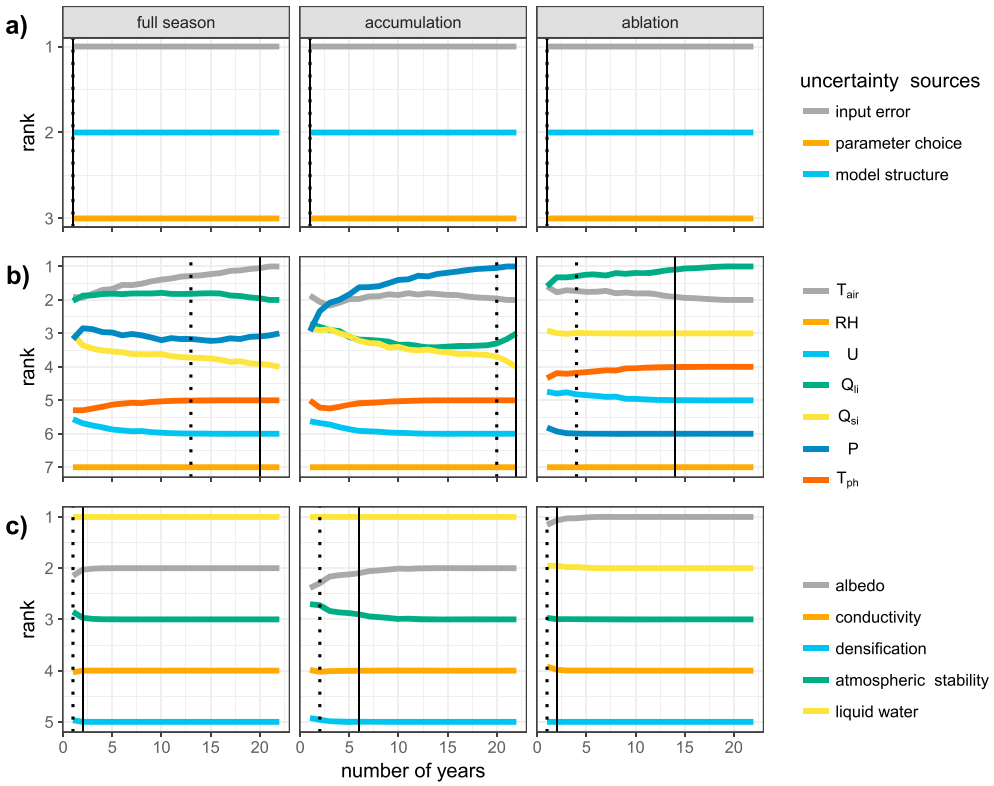

Figure 9. Ranking of the sensitivities for an increasing number of winter seasons. (a–c) The objectives (analyzed in sections 3.2.1, 3.2.2, and 3.2.3). Vertical solid and dotted lines indicate the number of years necessary to reproduce the true rank within ±0.1 and ±0.3, respectively.

means for each uncertainty group. The number of years that are necessary to obtain lower mean sensitivity errors than SEs of the mean over all winter seasons is indicated with dotted vertical lines. These thresholds are calculated for each uncertainty group but often result in the same value and hence are depicted only once or twice in each panel of Figure 8.

Comparing the three broader uncertainty classes input error, parameter choice, and model structure (Figure 8a) reveals that evaluating model predictions during only a few winter seasons (i.e., <10 years) does not suffice to estimate the mean sensitivity indices for all classes with reasonably accuracy (i.e., within the range of SE). In order to get a robust estimate of mean interannual accumulation and ablation sensitivities of individual forcing errors (Figure 8b), 9 to 10 and 10 years are necessary, respectively. Model structure sensitivities (Figure 8c) can also be resolved within SE precision using 9 to 10 winter seasons. Evaluating just during one single winter season can lead to S_{Ti} discrepancies up to 0.15 compared to estimations averaged over a longer time period.

Using the SE of the long-period mean as a key performance indicator might seem overly ambiguous. Absolute S_{Ti} errors appear indeed low for SA designs using one single winter season. Here the percentage of output variance explained by one uncertainty source can be determined with a mean accuracy of ±5.6%, ±5.6%, and ±3.2% for objectives (a), (b), and (c), respectively. However, absolute S_{Ti} errors should also be seen relative to the mean explanatory power of an uncertainty source (e.g., an absolute S_{Ti} error of ±0.056 is still significant when the uncertainty source only explains a fraction of 0.1 of the output variance). Furthermore, difference between the individual uncertainty sources are rather small in objectives (b) and (c); hence, not determining sensitivities within a high precision could lead also to a different ranking of importance.

In Figures 9a–9c, we present the resulting mean ranking of sensitivities when an increasing number of winter seasons are included in the SA. For objective (a) the differences between the three uncertainty groups input data error, model structure, and parameter choice are sufficiently large to obtain the correct ranking even for a 1-year analysis. Total-order sensitivity indices for individual forcing errors (objective b) are much closer together and show a much higher interannual variability (compare Figure 6). Hence, a longer

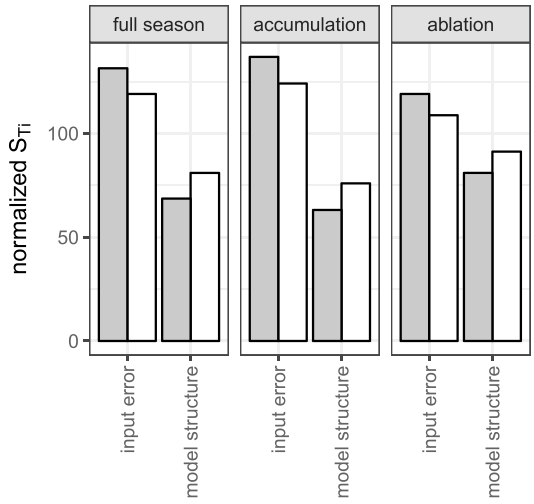

Figure 10. Mean interannual total-order sensitivity indices resulting from two different sensitivity analysis strategies normalized by its mean S_{Ti} value. (gray bars) Perturbing input error and model structure during the sensitivity analysis but use the default parameter set for all realizations. (white bars) Following the workflow presented in section 2.5 including perturbations of model parameters.

evaluation period is necessary in order to resolve these differences. A separation between all input error ranks (within ±0.3 of the true rank) requires 4 to 20 winter seasons. However, a clear separation (within ±0.1 of the true rank) can only be achieved after analyzing 20 and 14 years during the full season and the ablation period, respectively. Resolving all sensitivity ranks within an error range of ±0.1 during the accumulation period requires the whole time series. This indicates that at the presented site forcing recordings of 22 years might be not sufficient to obtain the true sensitivity order during the accumulation period. In fact, no individual year produced the same ranking of sensitivities to input errors compared to the ranking of the whole evaluation period during the full season. Mean rankings during the accumulation period could be reproduced in 1995 and during the ablation period in 1997, 2005, and 2006. Assessing the mean sensitivity ranks of individual process representations (objective c) shows that the ranks stabilize (with an error $<\pm 0.1$) after 2 years for the full season and the ablation period and 6 years are necessary to resolve the difference between the albedo and the atmospheric stability option during the accumulation period. Sensitivity ranks within an error range of ±0.3 can be obtained with a 1-year analysis. However, it is worth noting that even small errors in mean sensitivity ranks demonstrate that the true order could not be reproduced in all individual years. For example, during the full season analyses in 4 years do not result in the true sensitivity ranking and the correct ordering of sensitivities during the ablation period could not be reproduced in 7 years. In order to obtain the correct rankings in sensitivities for all three validation periods, an SA during one of the years 1995, 1997, 1998, 1999, 2000, 2009, 2010, 2014, and 2015 suffices. However, preselecting a single winter season for a SA is not a trivial task. Simple selection criteria based on meteorological conditions (as presented in Figure 1) might not be able identify representative years. Winter seasons with average conditions (within the interquartile range) in November–April air temperatures and maximum SWE (1992, 1994, 1995, 1998, 2002, 2004, 2008, and 2015) do include some but are not limited to the representative years listed above. Computational cost would greatly benefit from a preselection scheme able to identify winter seasons with representative sensitivity patterns. Hence, the development of such a scheme is certainly of interest for future studies.

3.4. Why Assess Sensitivities in the Face of other Uncertainties?

Assessing the impact of one or more uncertain parts in the modeling chain on output performance, while perturbing other uncertain parts is computationally more challenging than a classical SA, where just the factors of interests are changed. Our central argument is that to obtain more robust sensitivity estimates existing interaction effects should be included if possible. For example, many interactions might exist between the model structure choice and its parameters. Hence, in order to adequately assess the total impact of the model structure (e.g., compared to input data error) on output performance one should include the interaction effects of the parameter choice. Comparing the sensitivities with and without these interactions reveals this effect and is illustrated in Figure 10. S_{Ti} values are normalized by its mean. This conversion was necessary to ensure comparability between the two SA strategies. When omitting the interaction effect from parameter choice (in this example, just using the default parameters of the model combinations), input data error dominates the output variance more clearly. Including the varying parameter sets during the SA reduces the differences during the full season, the accumulation and ablation period alike.

Analogous considerations for the impact of individual forcing errors and model structures are shown in Figures 11 and 12. The estimation of input sensitivities are dependent upon the model and its parameters, and the effect a specific model option has on the simulation depends on the parameter values chosen and the forcing data used. In order to illustrate this effect, Figure 11 shows the difference between the SA strategy presented in section 2.5 (white bars) and an SA strategy including just one FSM configuration (FSM0: all options 0) and its default parameter set (gray bars). Values are again normalized by the mean. It can be seen that neglecting different model structures and parameter sets during the assessment of forcing error sensitivities can lead to very different findings. In this specific example discrepancies in total-order sensitivity estimates can be found especially for P, Q_{si}, Q_{li} and T_{air} errors. During the accumulation period PS_{Ti}

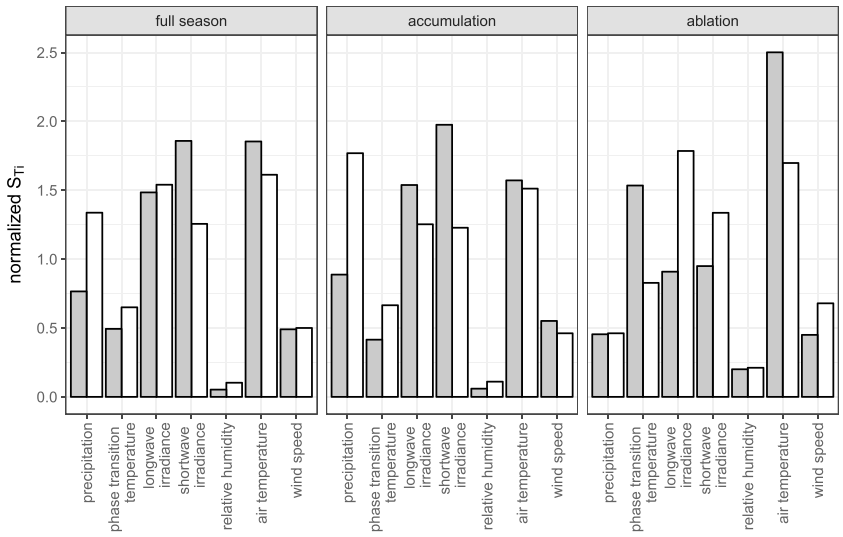

Figure 11. Mean interannual total-order sensitivity indices resulting from two different sensitivity analysis strategies normalized by its mean S_{Ti} value. (gray bars) Perturbing the shown parameters during the sensitivity analysis for just one snow model option (Factorial Snowpack Model configuration 0) and its default parametrization. (white bars) Following the workflow presented in section 2.5 including perturbations of model structure and parametrization.

values are underestimated, while during the ablation period Q_{si} and Q_{li} sensitivities are underestimated and the explanatory power of T_{air} overestimated.

In Figure 12, the SA results are shown as an example when no (artificial) input errors are introduced and snow model parameter values remain unchanged from its default. The two SA strategies result in different sensitivity patterns. While during the accumulation period the impact of the albedo and the liquid water option is overestimated, the impacts of the snow conductivity, densification, and stability correction option are underestimated resulting in a more unbalanced explanation of the output variance. During the ablation period the SA strategy with default parameter values overestimates, the impact of the snowpack liquid water transport scheme. When including input and parameter uncertainty, the albedo option is identified much

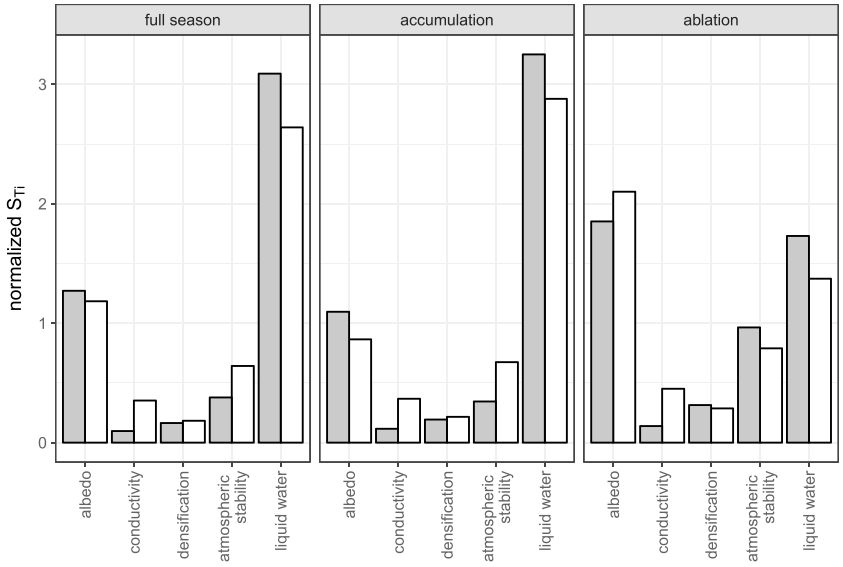

Figure 12. Mean interannual total-order sensitivity indices resulting from two different SA strategies normalized by its mean S_{Ti} value. (gray bars) Perturbing the shown parameters during the sensitivity analysis with no forcing error and the default model parametrization. (white bars) Following the workflow presented in section 2.5 including perturbations of forcing errors and model parametrization.

clearer as the dominant source of variance. Ignoring the uncertainties might lead to a distorted picture of the SA results.

4. Discussion

4.1. Sensitivity Indices

Input data quality is key for subsequent snow model simulations. The presented SA reveals the dominating effect of induced forcing errors on model performance throughout the accumulation and ablation period. However, the influence of the parameter choice on output variance is significantly lower than the influence of model structure throughout all performance criteria. Hence, we can accept only the first part of our initial hypothesis (ii; section 1).

Applying a constant forcing bias might not accurately mirror the error in input data one could expect from interpolation from surrounding station recordings. In fact, an error consisting of a bias, a dynamic effect (e.g., event based) and a random effect is assumed to be much more realistic. However, other studies have shown that the influence of a bias in the forcing data is more pronounced in the model output than a random error noise (Lapo et al., 2015; Raleigh et al., 2015). For the sake of simplicity, we assume that a bias in forcing data is able to encompass the true uncertainty range but want to acknowledge the uncertainty coming from this assumption as well.

The very low main effect values of the parameter choice are due to the sampling design of the SA. For every model combination, all parameters are sampled, ignoring whether the parameter is actually used in the model configuration or not. A varying number of parameters for the different model configurations forced us to employ this design, which leads to an unrealistic underestimation of the main effect but is not reflected in the total effect indices.

We found an increase in sensitivity of the model skill to all three analyzed uncertainty groups during the ablation period compared to the accumulation period. This increase in explanatory power is not unexpected, since different model realizations start to diverge during melting conditions. However, the total-order effect of the input data error still explains the largest proportion of the variance, given the increasing interactions with model structure and parameter choice during the ablation period.

While the impact of forcing error is critical for both the accumulation and the ablation period alike, the SA reveals that different input variables govern the variance (Figure 6). The results show a large interannual variability in output sensitivity to P errors compared to Q_{li}, Q_{si}, and T_{air} errors during the accumulation period. It is worth noting, again, that the sensitivity indices (i.e., the explanatory power) denote fractions of the output variance explained by the variable. Hence, these indices give the relative importance in the model system. In contrast to any other forcing variable, a negative correlation can be found when relating annual PS_{Ti} values and the performance spread (i.e., the variance) during the accumulation period (Pearson coefficient of correlation $-0.46, p = 0.025$). With increasing output variability, the sensitivity to P errors decreases. This indicates that a big part of the variability in PS_{Ti} values originates from the interannual variability of the other forcing variables.

The advection of heat by precipitation is neglected in FSM. Nonetheless, the results show that during the ablation period the errors of the precipitation phase becomes more important than the amount of precipitation, as the phase decision has considerable implications for the energy balance via various pathways. Rain falling on a cold snow surface (<0 °C) releases latent heat on refreezing, warms the snowpack, and potentially enables the onset of melt. In FSM, an increase in snowpack temperature toward melting conditions reduces the surface albedo, leading to enhanced energy input from shortwave radiation. A late snowfall, however, increases SWE and refreshes the albedo, limiting net shortwave radiation and consequently reduces the energy available for melt. Errors in the precipitation amount do not have such a strong impact on simulations during the ablation period, since precipitation tends to primarily fall as rain in spring. One might argue that not considering advection from precipitation might lead to an underestimation of the impact of the rainfall amount on model output; however, even in rain-on-snow environments advected energy was found to be minor, accounting for just 3% of the energy balance (Mazurkiewicz et al., 2008). In fact, a theoretical increase in rainfall of 5 mm/day at 5 °C would advect only about 1.2 W/m^2 into an already melting snowpack; however, it would release 19.3 W/m^2 of latent heat if the rain water freezes in the snow. Previous studies showed that a P bias is the most critical forcing error also for ablation rate predictions. The P errors introduced in this study are moderate compared to other studies, reducing the model sensitivity to

P by design. However, our results suggest that a large proportion of the *P* effect is in part a result from the phase determination, the effect of which was not resolved for in other SA.

The effect of both *RH* and *U* errors on ablation rates are diminished. In this study, we utilized a simple threshold air temperature to differentiate between snow and rainfall. In many environments, snowfall fraction was successfully linked to dew point or wet bulb temperature (e.g., Marks et al., 2013). Applying such an approximation method would certainly increase the total effect of *RH* errors, which show the smallest effect on model output performance in the presented SA. Due to the surrounding mountain ridges and the proximity to a forest stand (in the east and south), high wind speeds are not common at the Kühtai station. Consequently, the turbulent transfer of heat is limited and not a major source of energy.

We found the snowpack hydraulics option to have a large impact on snow mass simulations for both the accumulation and ablation period. We follow the argument from Essery et al. (2013), who links this finding to the release (or refreezing) of liquid water from winter surface melt and rainfall events. Due to its cumulative nature, these erroneous SWE predictions early in the winter season have a big effect in MAE.

4.2. Interaction Effects

The analysis showed that a considerable fraction of the output variance is explained by interaction effects. This finding is in line with our initial argument that assessing the sensitivities of such a model system should include a number of uncertainty sources for a better representativeness of the results. Including possible parameter values in the SA leads to a more robust estimation. For example, it seems obvious that the impact of switching on the prognostic liquid water transport option depends on the water holding capacity (i.e., the bucket size). With a reduction of this parameter (W_{irr}), the two hydraulic options behave more similarly. In fact, a value of $W_{\mathrm{irr}} = 0$ corresponds to hydraulic option 0. Hence, the impact of one specific process representation on output performance needs to be analyzed considering all meaningful parameter values. Otherwise, the analysis might tell us more about the differences between the specific model options used rather than about the underlying process. We show differences in the results with and without including multiple uncertainty sources during the SA (default parameter set, FSM configuration 0; Figures 10–12). Our results suggest that ignoring a source of uncertainty might not just affect sensitivity index values relative to each other (as shown in Figures 10 and 12) but can even change which variable is found to be the most influential on model output performance (Figure 11). These findings highlight the difficulty of extrapolating results from SA of a single model to other snow models and have direct implications for future model intercomparison studies as in the past these tended to focus primarily on model structures (e.g., Krinner et al., 2018; Lafaysse et al., 2017). The results show a considerable interannual variability of the interaction effect, indicating for years with high interactions that an improvement of knowledge (i.e., reduction of uncertainty) of one factor alone might not improve model results (Baroni & Tarantola, 2014).

4.3. Interannual Variability in Sensitivity Indices

Throughout the SA we observed a considerable interannual variability of calculated S_i and S_{Ti} values. Hence, statements about the mean explanatory power of a model component on output variance require a long evaluation period (at our study site). Long-term interannual variability in model sensitivity is yet to be examined for other environments. However, sites where the seasonal snowpack is not as variable (meaning the accumulation and ablation processes governing its evolution are similar between the years), are expected to show also lower interannual variability in the sensitivity pattern (e.g., very cold environments with no midwinter melt events). A robust estimate of sensitivity indices might be possible including fewer winter seasons in these environments.

4.4. Limitations

While the presented analysis shows a clear picture of model sensitivities and illustrates the importance of including various uncertainty sources within a SA, it is important to keep in mind the following caveats. In this study, the uncertainties in snow modeling are quantified following previous studies and published measurements (section 2.5.1). However, one might argue that the selection of these uncertainty distributions is rather subjective if not arbitrary and might not reflect the true uncertainty inherent to the system. It is clear that the results obtained from SA are dependent upon the error and parameter ranges samples are drawn from. For example, Raleigh et al. (2015) performed SA to input errors for different scenarios differing in their representation of *P* errors. For these different error scenarios they found not only very different sensitivities for *P* but also for all other forcing variables. We also acknowledge that different snowpack process descriptions might yield different SA results and that the applied model structures might not reflect

the variance coming from a larger ensemble. A robust quantification of the uncertainty encountered when using spatially interpolated forcing data and various snow models is not trivial. In this study, input errors and model parameter ranges are based on surface observations, representing the only meaningful reference.

As is the case for most climate stations (Raleigh et al., 2016), the Kühtai station is lacking observations of incoming longwave irradiance. Constructing this energy flux from other meteorological variables is therefore a common surrogate. In this study, the longwave approximation was performed off-line the SA, hence treating longwave irradiance the same way as all the measured forcing variables. As a result no interaction between the longwave approximation and other errors are considered. However, the true uncertainty associated with the approximated Q_{li} might be higher, as Q_{li} errors are directly linked to Q_{si}, T_{air}, and RH errors. Furthermore, the inclusion of longwave emission from surrounding slopes results in indirect linkages between calculated Q_{li} and all other uncertainty sources (section 2.2). Hence, including this interplay (i.e., compute Q_{li} online) increases total-order sensitivity indices for longwave irradiance, due to the increase of interaction effects.

The ensemble spread resulting from all (objective a) presented model realizations (Figure 2) was shown not to encompass observations in some years, implying that the overall uncertainty was not captured in the analysis. One possible explanation is the low mean of the P error distribution. As this error is based on the difference of two measurement techniques that are potentially both subject to wind-induced undercatch, the resulting error distribution might suffer from a bias. Previous studies suggested that validation data quality plays an important role and might limit snow model performance drastically (Magnusson et al., 2015). Spatially distributed snow models are also often evaluated using automated measurements at the point scale (e.g., sonic ranger or snow pillow recordings). However, due to topographic and microclimatic effects on accumulation, redistribution, and ablation processes, the snow cover is spatially heterogeneous even at very small scales (López-Moreno et al., 2011), introducing an uncertainty in validation data (i.e., in the subgrid scale). Evaluating the model performance exclusively on SWE observations limits the informative value of the results and using multiple working hypotheses has been advocated instead (Clark et al., 2011). Lapo et al. (2015) found many cases where perturbed forcing irradiances were not manifested in SWE simulations but notably in snow surface temperature. In the presented study, model skill was restricted to snow pillow recordings and this equifinality accepted, as (i) surface temperature observations are lacking and available snow temperatures profiles proved to be rather erroneous at the site, and (ii) linking the sensitivity values directly to the energy balance components was not within the scope of this study rather than introducing a methodology. Therefore, sensitivity indices were also just computed for two parts of the snow season (accumulation and ablation period) and not at shorter time scales (e.g., event basis).

5. Conclusions

Several snow model intercomparison studies were not able to fully link model performance to model structure (Essery et al., 2013; Magnusson et al., 2015). No single best model could be identified, partly due to nonlinearity of models, the degree of interactions between virtually all parts of the system, and the resulting compensation effects. In this study, we presented a workflow able to include multiple uncertainty sources in a SA, while preserving interpretability and computational feasibility. Following this workflow, we were able to (a) compare the impact of forcing data error, model structure, and parameter choice on snow model performance; (b) compare the sensitivity of model performance to forcing data errors for a wide range of model structures and parameter sets; and (c) assess the impact of individual model options in the face of parameter and forcing data uncertainty.

A key consideration controlling model skill proved to be the input data quality. This was true during both the accumulation and ablation periods. Over the whole analyzed period model skill variance was governed in the order of parameter choice < model structure < input data error. However, for the prediction of melt rates (negative SWE changes) and the timing of maximum SWE, sensitivities to input error and model structure are comparable in size. The model's ability to reproduce mean seasonal melt rates is more sensitive to the model structure than to forcing errors. While errors in precipitation amount, air temperature, and radiative forcings dominate the variance during the accumulation period, the impacts of precipitation phase, longwave and shortwave irradiance, wind speed, and air temperature become larger than the impact of precipitation amount during the ablation period. The analysis confirmed the importance of the albedo representation found in previous studies, especially during the ablation (Blöschl, 1991; Essery et al., 2013;

Magnusson et al., 2015). A strong sensitivity of model skill to the snowpack hydraulics option was revealed throughout the full winter season. As noted by Essery et al. (2013), at sites where surface melt is common during the winter, at least a simple liquid water transport scheme is required in order to allow for storage and refreezing.

We argue that including multiple uncertainty sources while assessing the impact of individual components in the model chain leads to more robust results, given the high degree of interactions in the system. This hypothesis is tested by comparing SA designs with and without interactions, resulting in very different findings. For the specific cases we presented in section 3.4, this leads not just to differences in the relative importance of individual variables (Figures 10 and 12) but potentially also to very different conclusions about which variable is most important in explaining the variance (i.e., has the largest impact on model performance; Figure 11). Therefore, extrapolating results from single model sensitivity analyses to different snow model structures is questionable at best. Future model intercomparison studies need to take these interaction effects into account.

This study also demonstrated a considerable interannual variability in computed sensitivity values and their resulting rankings. We advocate the evaluation of snow models over multiple years (>10), when conclusions about average, site-specific sensitivities are to be drawn from the analysis. The evaluation time required in order to replicate the correct sensitivity ranking of all considered elements within a SA is dependent upon the objective. While it is sufficient to resolve differences between forcing errors, model structure, and parameter choice during a 1-year analysis at the presented site, up to 6 and 22 years are required when ranking sensitivities to individual model structures and forcing errors, respectively.

We investigated the impact of various choices a modeler typically has to make when simulating seasonal snow cover at a given point but did not show which settings perform better or worse than others. Transferring this idea of simultaneously varying model structures and parameter sets into spatially distributed modeling, where forcing uncertainties are difficult to control or quantify, poses an interesting computational challenge. Comparing resulting model skills at point and catchment scale for a suite of forcing error scenarios might allow a deeper understanding of robust model structures. A manuscript addressing these issues is in preparation.

Acknowledgments
We thank Mark Raleigh, Adam Winstral, and an anonymous referee for their constructive comments, which helped to sharpen the scope of this study. This work was funded by the EUREGIO research project CRYOMON-SciPro FIPN000100. Development of FSM is supported by NERC grant NE/P011926/1. Publication of the manuscript is gratefully supported by the University of Innsbruck. The data set of all modeled time series is available from the Zenodo repository (doi:10.5281/zenodo.1466038).

References

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*, W01403. https://doi.org/10.1029/2005WR004745

Baroni, G., & Tarantola, S. (2014). A General Probabilistic Framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. *Environmental Modelling & Software*, *51*, 26–34. https://doi.org/10.1016/J.ENVSOFT.2013.09.022

Bartelt, P., & Lehning, M. (2002). A physical SNOWPACK model for the Swiss avalanche warning: Part I: Numerical model. *Cold Regions Science and Technology*, *35*(3), 123–145. https://doi.org/10.1016/S0165-232X(02)00074-5

Blöschl, G. (1991). The influence of uncertainty in air temperature and albedo on snowmelt. *Nordic hydrology*, *22*(2), 95–108.

Blöschl, G., Gutknecht, D., & Kirnbauer, R. (1991). Distributed snowmelt simulations in an alpine catchment: 2. Parameter study and model predictions. *Water Resources Research*, *27*(12), 3181–3188. https://doi.org/10.1029/91WR02251

Blöschl, G., & Kirnbauer, R. (1991). Point snowmelt models with different degrees of complexity—Internal processes. *Journal of Hydrology*, *129*(1-4), 127–147. https://doi.org/10.1016/0022-1694(91)90048-M

Blöschl, G., & Kirnbauer, R. (1992). An analysis of snowcover patterns in a small alpine catchment. *Hydrological Processes*, *6*(April 1991), 99–109.

Blöschl, G., Kirnbauer, R., & Gutknecht, D. (1991). Distributed snowmelt simulations in an alpine catchment: 1. Model evaluation on the basis of snow cover patterns. *Water Resources Research*, *27*(12), 3171–3179. https://doi.org/10.1029/91WR02250

Brock, B. W., Willis, I. C., & Sharp, M. J. (2006). Measurement and parameterization of aerodynamic roughness length variations at Haut Glacier d'Arolla, Switzerland. *Journal of Glaciology*, *52*(177), 281–297. https://doi.org/10.3189/172756506781828746

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Cline, D., Pomeroy, J. W., & Gray, D. M. (1998). Snowcover accumulation, relocation and management. *Arctic and Alpine Research*, *30*(3), 314. https://doi.org/10.2307/1551979

Corripio, J. G. (2002). Modelling the energy balance of high altitude glacierised basins in the Central Andes (Ph.D. thesis), University of Edingburgh.

Daniel Moore, R. (1983). On the use of bulk aerodynamic formulae over melting snow. *Hydrology Research*, *14*(4), 193–206. https://doi.org/10.2166/nh.1983.0016

Douville, H., Royer, J., & Mahfouf, J. (1995). A new snow parameterization for the Meteo-France climate model. *Climate Dynamics*, *35*, 21–35. https://doi.org/10.1007/BF00208760

Essery, R. (2015). A Factorial Snowpack Model (FSM 1.0). *Geoscientific Model Development*, *8*(12), 3867–3876. https://doi.org/10.5194/gmd-8-3867-2015

Essery, R., Morin, S., Lejeune, Y., & Ménard, C. (2013). A comparison of 1701 snow models using observations from an alpine site. *Advances in Water Resources*, *55*, 131–148. https://doi.org/10.1016/j.advwatres.2012.07.013

Etchevers, P., Martin, E., Brown, R., Fierz, C., Lejeune, Y., Bazile, E., et al. (2004). Validation of the energy budget of an alpine snowpack simulated by several snow models (Snow MIP project). *Annals of Glaciology*, *38*, 150–158. https://doi.org/10.3189/172756404781814825

Greuell, W., Knap, W. H., & Smeets, P. C. (1997). Elevational changes in meteorological variables along a midlatitude glacier during summer. *Journal of Geophysical Research*, *102*(D22), 25,941–25,954. https://doi.org/10.1029/97JD02083

Gromke, C., Manes, C., Walter, B., Lehning, M., & Guala, M. (2011). Aerodynamic roughness length of fresh snow. *Boundary-Layer Meteorology*, *141*(1), 21–34. https://doi.org/10.1007/s10546-011-9623-3

Günther, D., Hanzer, F., Marke, T., Essery, R., & Strasser, U. (2017). Sensitivitätsanalyse energiebilanzbasierter schneemodelle. 2. Workshop zur Alpinen Hydrologie. Hydrologische Prozesse im Hochgebirge im Wandel der Zeit, 15.-17.11, 2017, Obergurgl, Austria, Deutsche Hydrologische Gesellschaft, Österreichigsche Hydrologische Gesellschaft.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1-2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Harder, P., & Pomeroy, J. W. (2013). Estimating precipitation phase using a psychrometric energy balance method. *Hydrological Processes*, *27*(13), 1901–1914. https://doi.org/10.1002/hyp.9799

Harder, P., & Pomeroy, J. W. (2014). Hydrological model uncertainty due to precipitation-phase partitioning methods. *Hydrological Processes*, *28*(14), 4311–4327. https://doi.org/10.1002/hyp.10214

Helgason, W., & Pomeroy, J. W. (2011). Characteristics of the near-surface boundary layer within a mountain valley during winter. *Journal of Applied Meteorology and Climatology*, *51*(3), 583–597. https://doi.org/10.1175/JAMC-D-11-058.1

Jansen, M. J. (1999). Analysis of variance designs for model output. *Computer Physics Communications*, *117*(1-2), 35–43. https://doi.org/10.1016/S0010-4655(98)00154-4

Jennings, K. S., Winchell, T. S., Lineh, B. N., & Molotch, P. (2018). Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere. *Nature Communications*, *9*(1), 1148. https://doi.org/10.1038/s41467-018-03629-7

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1 Theory. *Water Resources Research*, *42*, W03407. https://doi.org/10.1029/2005WR004368

Kirnbauer, R., & Blöschl, G. (1990). A lysimetric snow pillow station at Kühtai/Tyrol (Vol. 193, pp. 463–470).

Klok, E. J., & Oerlemans, J. (2002). Model study of the spatial distribution of the energy and mass balance of Morteratschgletscher, Switzerland. *Journal of Glaciology*, *48*(163), 505–518.

Krajči, P., Kirnbauer, R., Parajka, J., Schöber, J., & Blöschl, G. (2017). The Kühtai data set: 25 years of lysimetric, snow pillow, and meteorological measurements. *Water Resources Research*, *53*, 5158–5165. https://doi.org/10.1002/2017WR020445

Krinner, G., Derksen, C., Essery, R., Flanner, M., Hagemann, S., Clark, M., et al. (2018). ESM-SnowMIP: Assessing snow models and quantifying snow-related climate feedbacks. *Geoscientific Model Development*, *11*(12), 5027–5049. https://doi.org/10.5194/gmd-11-5027-2018

Lafaysse, M., Cluzet, B., Dumont, M., Lejeune, Y., Vionnet, V., & Morin, S. (2017). A multiphysical ensemble system of numerical snow modelling. *The Cryosphere*, *11*, 1173–1198. https://doi.org/10.5194/tc-11-1173-2017

Lapo, K. E., Hinkelman, L. M., Raleigh, M. S., & Lundquist, J. D. (2015). Impact of errors in the downwelling irradiances on simulations of snow water equivalent, snow surface temperature, and the snow energy balance. *Water Resources Research*, *51*, 1649–1670. https://doi.org/10.1002/2014WR016259

Liston, G. E., & Elder, K. (2006). A meteorological distribution system for high-resolution terrestrial modeling (MicroMet). *Journal of Hydrometeorology*, *7*(2), 217–234. https://doi.org/10.1175/JHM486.1

López-Moreno, J. I., Fassnacht, S. R., Beguería, S., & Latron, J. B. P. (2011). Variability of snow depth at the plot scale: Implications for mean depth estimation and sampling strategies. *The Cryosphere*, *5*, 617–629. https://doi.org/10.5194/tc-5-617-2011

Louis, J.-F., Tiedtke, M., & Geleyn, J.-F. (1982). A short history of the PBL parameterization at ECMWF. Workshop on Planetary Boundary Layer parameterization, 25-27 November 1981 (pp. 59-79). ECMWF, Shinfield Park, Reading.

Magnusson, J., Wever, N., Essery, R., Helbig, N., Winstral, A., & Jonas, T. (2015). Evaluating snow models with varying process representations for hydrological applications. *Water Resources Research*, *51*, 2707–2723. https://doi.org/10.1002/2014WR016498

Marks, D., Kimball, J., Tingey, D., & Link, T. (1998). The sensitivity of snowmelt processes to climate conditions and forest during rain on snow (SNOBAL).pdf. *Hydrological Processes*, *1587*(March), 1569–1587.

Marks, D., Winstral, A., Reba, M., Pomeroy, J., & Kumar, M. (2013). An evaluation of methods for determining during-storm precipitation phase and the rain/snow transition elevation at the surface in a mountain basin. *Advances in Water Resources*, *55*, 98–110. https://doi.org/10.1016/j.advwatres.2012.11.012

Mazurkiewicz, A. B., Callery, D. G., & McDonnell, J. J. (2008). Assessing the controls of the snow energy balance and water available for runoff in a rain-on-snow environment. *Journal of Hydrology*, *354*(1-4), 1–14. https://doi.org/10.1016/j.jhydrol.2007.12.027

Mosier, T. M., Hill, D. F., & Sharp, K. V. (2016). How much cryosphere model complexity is just right? Exploration using the conceptual cryosphere hydrology framework. *The Cryosphere*, *10*(5), 2147–2171. https://doi.org/10.5194/tc-10-2147-2016

Parajka, J. (2017). The Kühtai dataset: 25 years of lysimetric, snow pillow and meteorological measurements [Data set]. Zenodo, https://doi.org/10.5281/zenodo.556110

Pomeroy, J. W., & Brun, E. (2001). *Physical properties of snow* (pp. 45–126). Berlin/Heidelberg: Springer Reference, Springer-Verlag. https://doi.org/10.1007/SpringerReference_225906

Pomeroy, J. W., Fang, X., & Marks, D. G. (2016). The cold rain-on-snow event of June 2013 in the Canadian Rockies—Characteristics and diagnosis. *Hydrological Processes*, *30*(17), 2899–2914. https://doi.org/10.1002/hyp.10905

Raleigh, M. S., Livneh, B., Lapo, K., & Lundquist, J. D. (2016). How does availability of meteorological forcing data impact physically based snowpack simulations? *Journal of Hydrometeorology*, *17*(1), 99–120. https://doi.org/10.1175/JHM-D-14-0235.1

Raleigh, M. S., Lundquist, J. D., & Clark, M. P. (2015). Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework. *Hydrology and Earth System Sciences*, *19*(7), 3153–3179. https://doi.org/10.5194/hess-19-3153-2015

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181*(2), 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Saltelli, A., Ratto, M., Tarantola, S., & Campolongo, F. (2006). Sensitivity analysis practices: Strategies for model-based inference. *Reliability Engineering and System Safety*, *91*(10-11), 1109–1125. https://doi.org/10.1016/j.ress.2005.11.014

Sauter, T., & Obleitner, F. (2015). Assessing the uncertainty of glacier mass-balance simulations in the European Arctic based on variance decomposition. *Geoscientific Model Development*, *8*(12), 3911–3928. https://doi.org/10.5194/gmd-8-3911-2015

Singh, P., & Singh, V. (2001). *Snow and glacier hydrology* (6th ed., 105 pp.). Dordrecht: Kluwer Academic Publishers.
Sobol, I. M. (1993). Sensitivity analysis for nonlinear mathematical models. *Mathematical Models and Computer Exp*, *1*(4), 407–414. https://doi.org/10.18287/0134-2452-2015-39-4-459-461
Strasser, U. (2008). Modelling of the mountain snow cover in the Berchtesgaden National Park, Tech. rep.
Strasser, U., Corripio, J., Pellicciotti, F., Burlando, P., Brock, B., & Funk, M. (2004). Spatial and temporal variability of meteorological variables at Haut Glacier d'Arolla (Switzerland) during the ablation season 2001: Measurements and simulations. *Journal of Geophysical Research*, *109*, D03103. https://doi.org/10.1029/2003JD003973
Strasser, U., & Marke, T. (2010). ESCIMO.spread—A spreadsheet-based point snow surface energy balance model to calculate hourly snow water equivalent and melt rates for historical and changing climate conditions. *Geoscientific Model Development*, *3*(2), 643–652. https://doi.org/10.5194/gmd-3-643-2010
Sturm, M., Perovich, D. K., & Holmgren, J. (2002). Thermal conductivity and heat transfer through the snow on the ice of the Beaufort Sea. *Journal of Geophysical Research*, *107*(C10), 8043. https://doi.org/10.1029/2000JC000409
Tarboton, D., & Luce, C. (1996). Utah energy balance snow accumulation and melt model (UEB). Computer model technical description and users guide, (December) 1–64.
Vionnet, V., Brun, E., Morin, S., Boone, A., Faroux, S., Le Moigne, P., et al. (2012). The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2. *Geoscientific Model Development*, *5*(3), 773–791. https://doi.org/10.5194/gmd-5-773-2012
Wang, L., MacKay, M., Brown, R., Bartlett, P., Harvey, R., & Langlois, A. (2013). Application of satellite data for evaluating the cold climate performance of the Canadian Regional Climate Model over Québec, Canada. *Journal of Hydrometeorology*, *15*(2), 614–630. https://doi.org/10.1175/JHM-D-13-086.1
Winstral, A., Marks, D., & Gurney, R. (2013). Simulating wind-affected snow accumulations at catchment to basin scales. *Advances in Water Resources*, *55*, 64–79. https://doi.org/10.1016/j.advwatres.2012.08.011
Ye, H., Cohen, J., & Rawlins, M. (2013). Discrimination of solid from liquid precipitation over Northern Eurasia using surface atmospheric conditions. *Journal of Hydrometeorology*, *14*(4), 1345–1355. https://doi.org/10.1175/JHM-D-12-0164.1